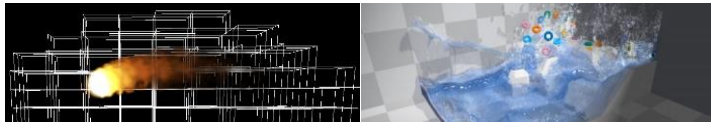
# CIRRUS:
# AI & DATA SCIENCE TOOLS, SOFTWARE AND TECHNOLOGIES

Paul Graham | Senior Solutions Architect | NVIDIA

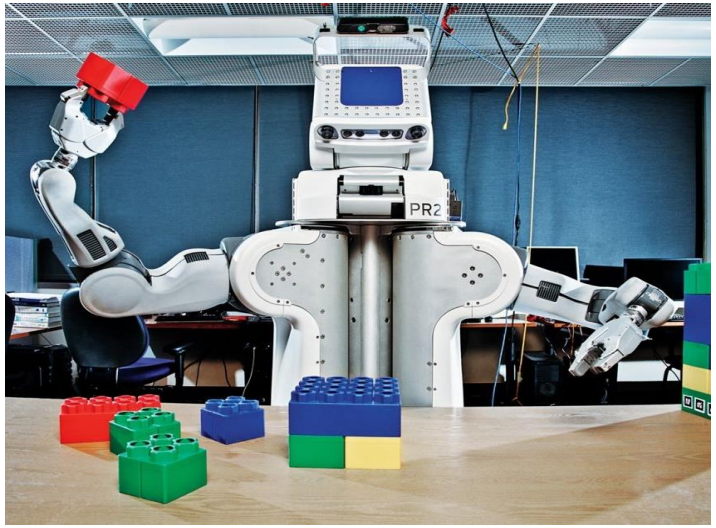
# NVIDIA



GPU Computing

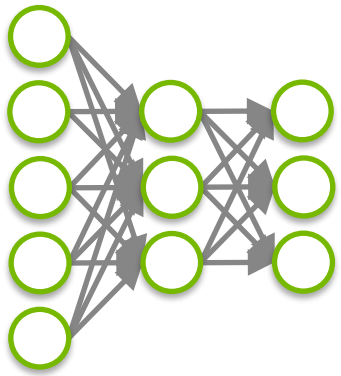


Computer Graphics




Artificial Intelligence

# THE BIG BANG IN MACHINE LEARNING

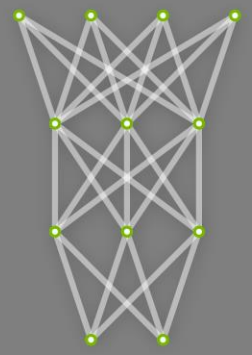


DNN


GPU

# DEEP LEARNING

**Untrained**
Neural Network
Model



# DEEP LEARNING

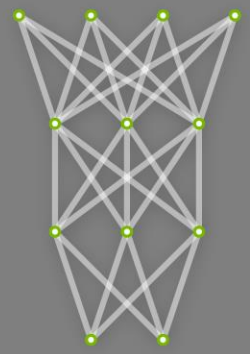
## TRAINING

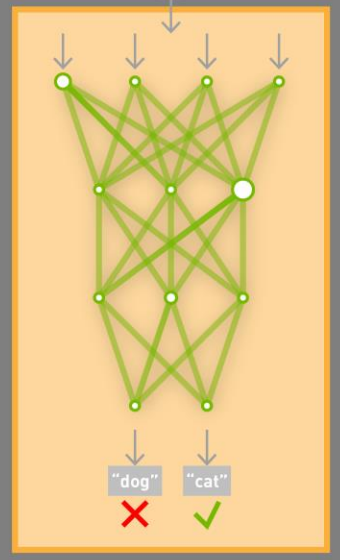
Learning a new capability
from existing data

**Untrained**
Neural Network
Model

Deep Learning
**Framework**

TRAINING
DATASET





# DEEP LEARNING

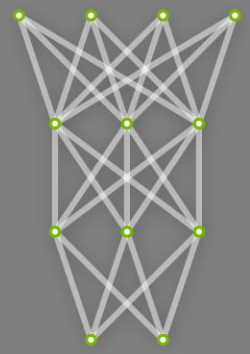
## TRAINING

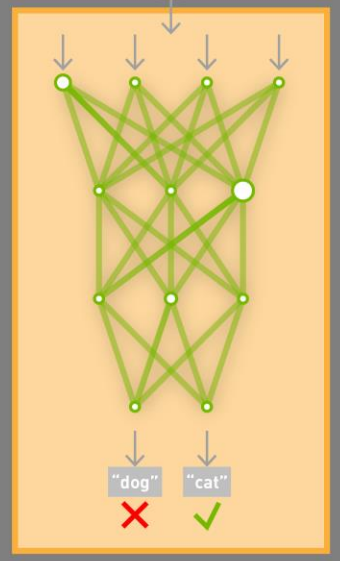
Learning a new capability
from existing data

**Untrained**
Neural Network
Model

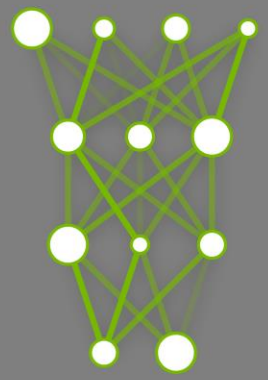
Deep Learning
**Framework**

TRAINING
DATASET




**Trained Model**
New Capability



# DEEP LEARNING

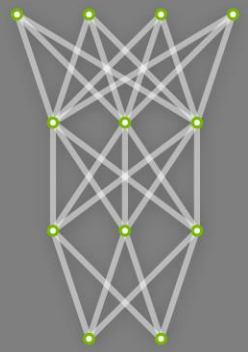
## TRAINING

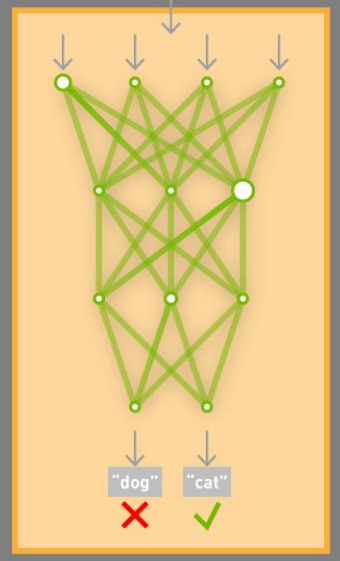
Learning a new capability
from existing data

**Untrained**
Neural Network
Model

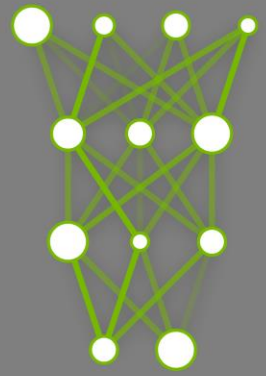
Deep Learning
**Framework**

TRAINING
DATASET



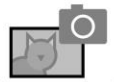

**Trained Model**
New Capability



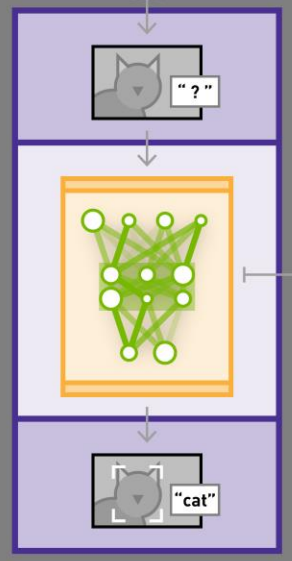
## INFERENCE

Applying this capability
to new data

NEW
DATA

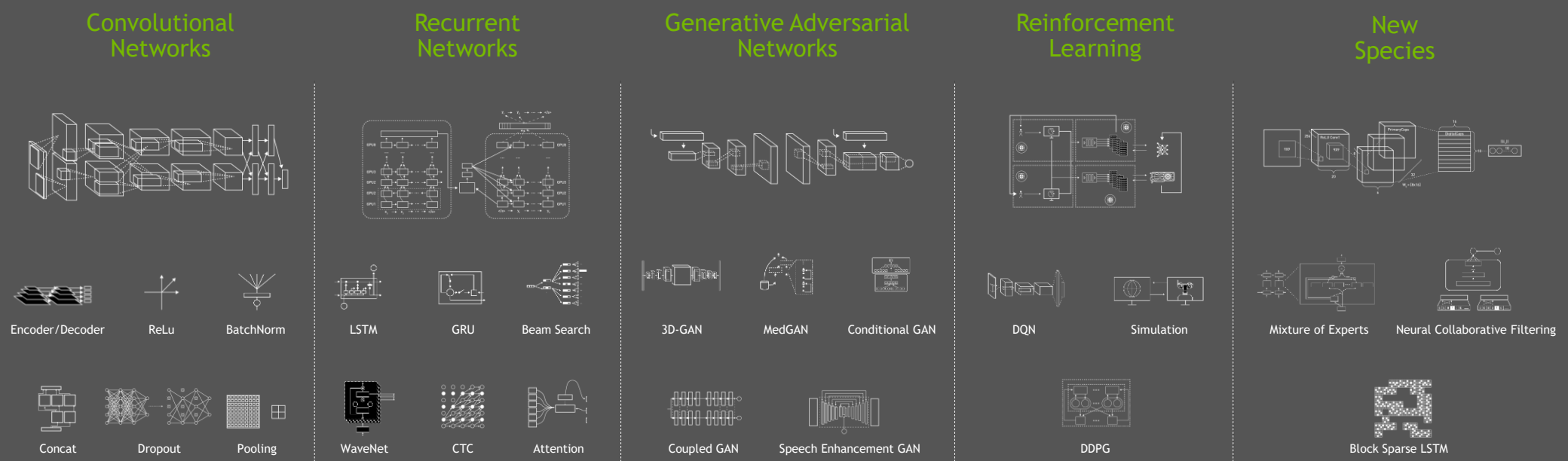


**App or Service**
Featuring Capability



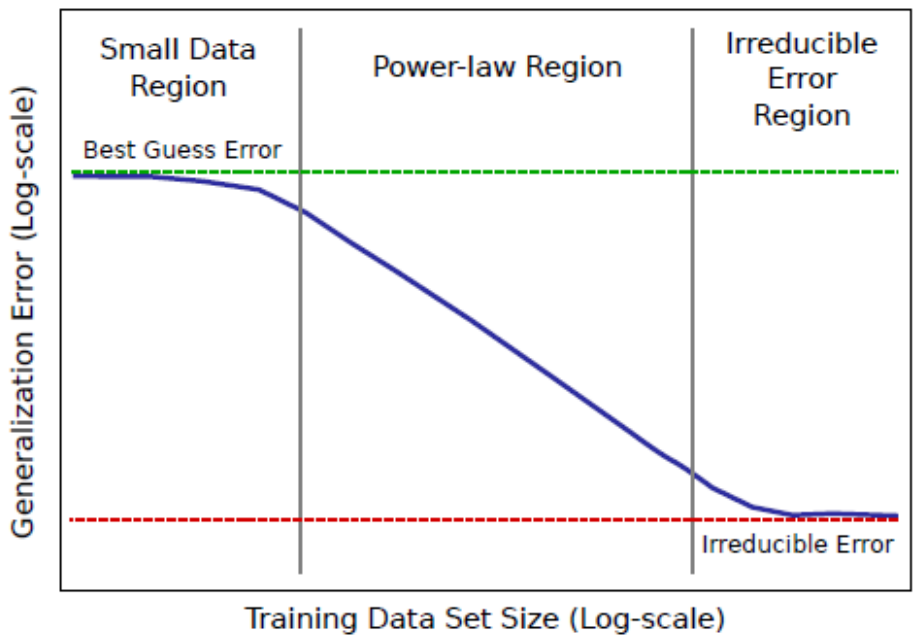

**Trained Model**
Optimized for
Performance

# CAMBRIAN EXPLOSION



# EXPLODING DATASETS

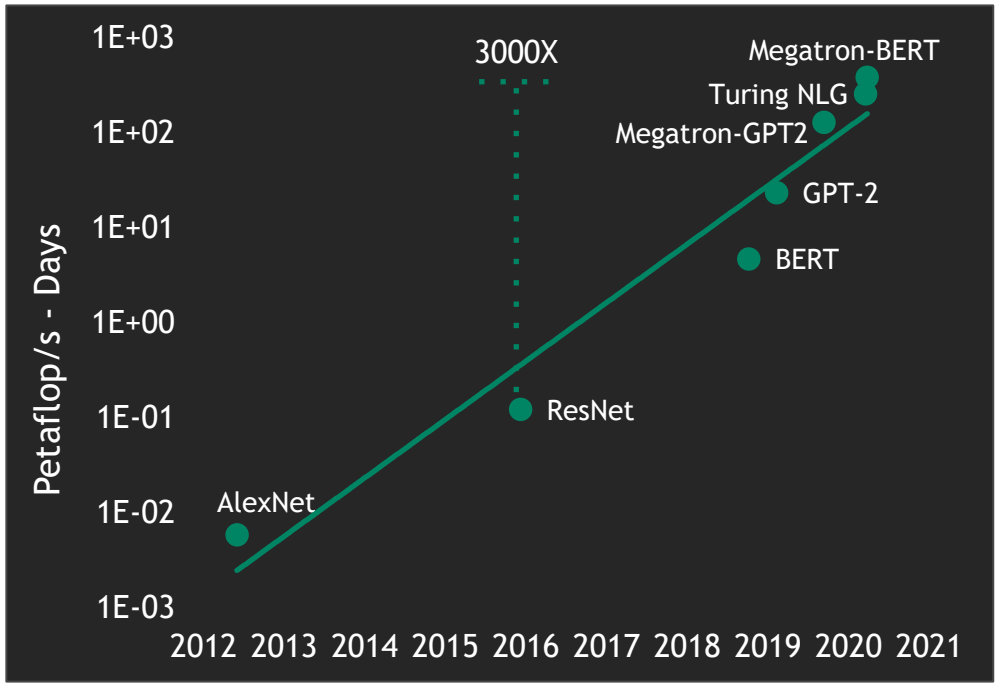
## Logarithmic relationship between the dataset size and accuracy




Hestness, J., Narang, S., Ardalani, N., Diamos, G., Jun, H., Kianinejad, H., ... & Zhou, Y. (2017). Deep Learning Scaling is Predictable, Empirically. arXiv preprint arXiv:1712.00409.

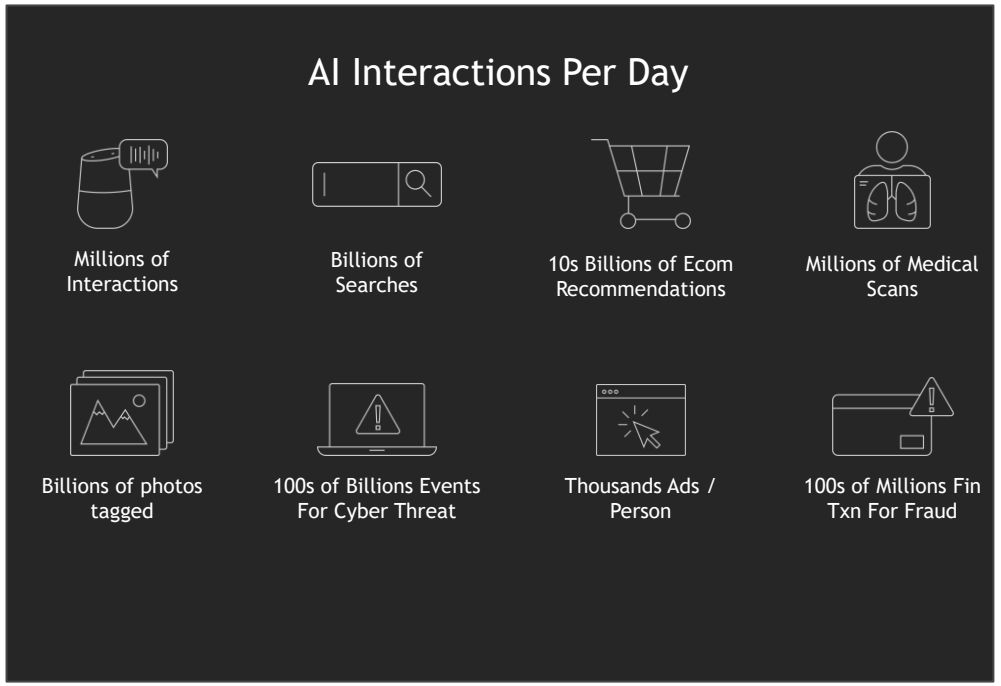
# CHALLENGES: ACCELERATING BIG AND SMALL

## AI Advances Demand Exponentially Higher Compute




3000X Higher Compute Required to Train Largest Models Since Volta

## AI Applications Demand Distributed Pervasive Acceleration

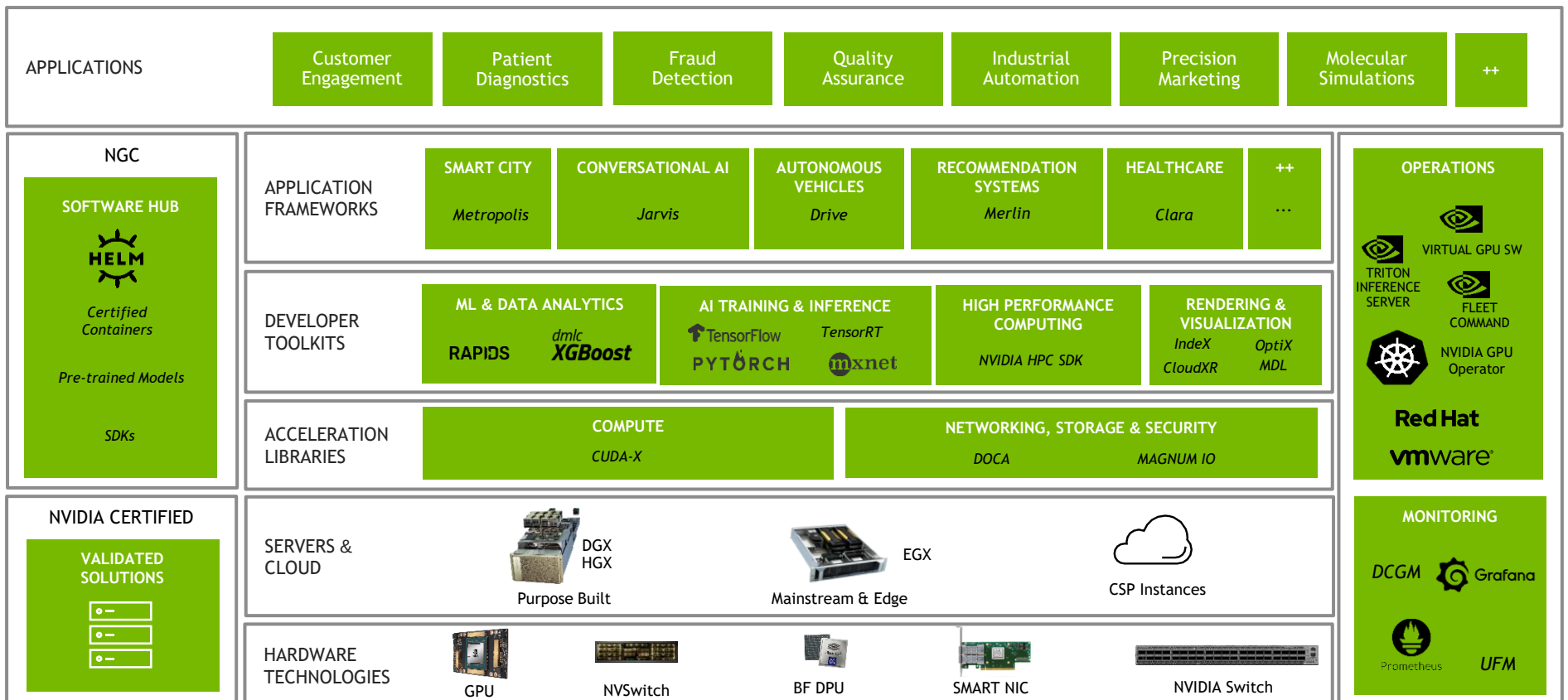



Every AI Powered Interaction Needs Varying Amount of Compute

Source: OpenAI, NVIDIA

# SOFTWARE & FRAMEWORKS

# NVIDIA PLATFORM



# CUDNN

## Accelerating DL primitives

### Key Features

- Tensor Core acceleration for all popular convolutions
- Optimized kernels for computer vision and speech models
- Supports FP32, FP16, and TF32 floating point formats and INT8, and UINT8 integer formats
- Arbitrary dimension ordering, striding, and sub-regions for 4d tensors means easy integration into any neural net implementation
- Speed up fused operations on any CNN architecture

### cuDNN 8 highlights include:

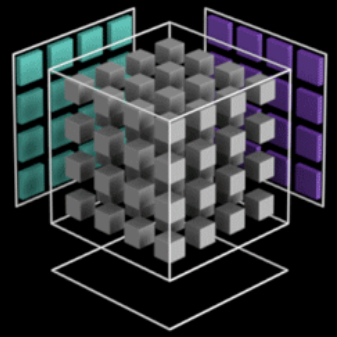
- Tuned for peak performance on NVIDIA A100 GPUs including new TensorFloat-32, FP16, and FP32
- Redesigned low-level API provides direct access to cuDNN kernels for greater control and performance tuning
- Backward compatibility layer maintains support for cuDNN 7.x letting developers manage their transition to the new cuDNN 8 API
- New optimizations for computer vision, speech, and language understanding networks
- Fuse operators to accelerate convolutional neural networks with a new API

# TENSOR CORES

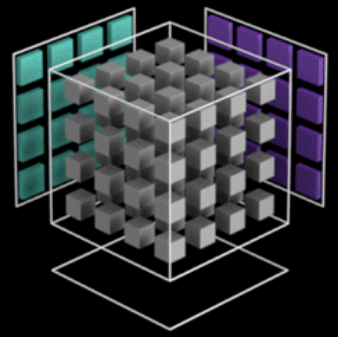
## Hardware for Matrix Multiply and Accumulate operations

- Introduced in the V100
- Perform several MMA calcs per clock cycle
  - FP32 in, FP32 out (accumulate)
  - FP16 multiply
- Turing added int8, int4 calculations
- Ampere
  - Full FP64 MMA
  - Bfloat16, Tensor Float 32

PASCAL



VOLTA TENSOR CORES



# CUTLASS – TENSOR CORE PROGRAMMING MODEL

## Warp-Level GEMM and Reusable Components for Linear Algebra Kernels in CUDA

### CUTLASS 2.3

GEMMs targeting structured Sparse Tensor Cores in NVIDIA Ampere Architecture GPUs

Fast SGEMM kernels targeting GeForce RTX 30-series CUDA Cores

### CUTLASS 2.2

Optimal performance on NVIDIA Ampere microarchitecture

New floating-point types: `nv_bfloat16, TF32, double`

Deep software pipelines with async memcopy

### CUTLASS 2.0

Significant refactoring using modern C++11 programming

```
using Mma = cutlass::gemm::warp::DefaultMmaTensorOp<
  GemmShape<64, 64, 16>,
  half_t, LayoutA, // GEMM A operand
  half_t, LayoutB, // GEMM B operand
  float, RowMajor // GEMM C operand
>;

__shared__ ElementA smem_buffer_A[Mma::Shape::kM * GemmK];
__shared__ ElementB smem_buffer_B[Mma::Shape::kN * GemmK];

// Construct iterators into SMEM tiles
Mma::IteratorA iter_A({smem_buffer_A, lda}, thread_id);
Mma::IteratorB iter_B({smem_buffer_B, ldb}, thread_id);

Mma::FragmentA frag_A;
Mma::FragmentB frag_B;
Mma::FragmentC accum;

Mma mma;

accum.clear();

#pragma unroll 1
for (int k = 0; k < GemmK; k += Mma::Shape::kK) {

  iter_A.load(frag_A); // Load fragments from A and B matrices
  iter_B.load(frag_B);

  ++iter_A; ++iter_B;  // Advance along GEMM K to next tile in A
                       // and B matrices

                       // Compute matrix product
  mma(accum, frag_A, frag_B, accum);
}
```

# AMP

Automatic Mixed Precision in DL training

| FP32 | FP16 with Tensor Cores* |
| --- | --- |
| 1x compute throughput | 8X compute throughput |
| 1x memory throughput | 2X memory throughput |
| 1x memory storage | 1/2X memory storage |

*figures are for V100

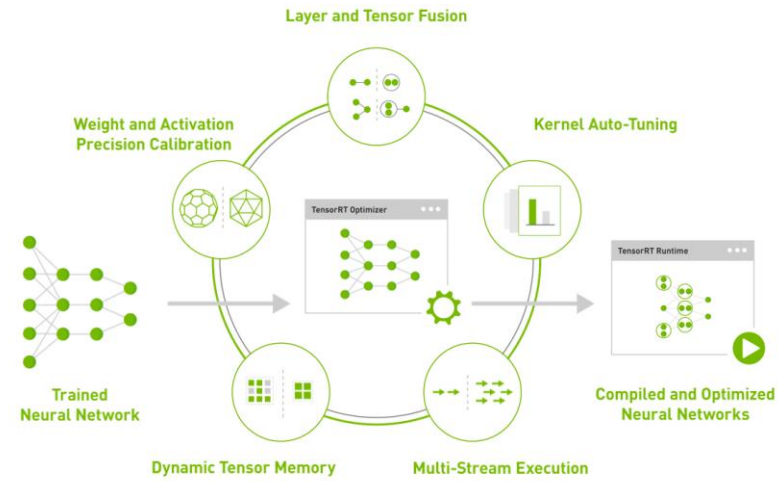
# MULTIGPU

- DL Frameworks on NGC already have multiGPU, multi-node support built in
- For programmers, various technologies
  - NVLink / NVSwitch
  - Horovod
  - NVSHMEM
  - NCCL
  - GPUDirect
  - Analysis: Nsight - DLprof, PyProf

# TensorRT

## Accelerating Inference

### SDK for High-Performance Deep Learning Inference




Optimize and Deploy neural networks in production

Maximize throughput for latency-critical apps with compiler & runtime

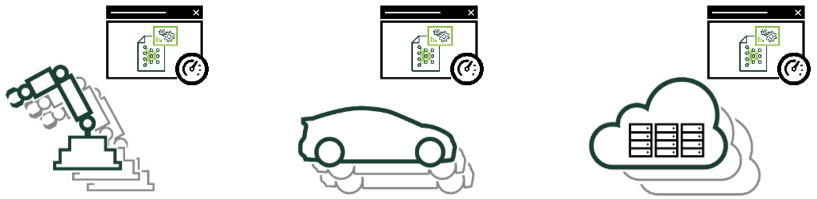
Deploy responsive and memory efficient apps

FP32, TF32, BFLOAT16, FP16 & INT8

Optimize every network including CNNs, RNNs and Transformers

Accelerate every framework - ONNX support, TensorFlow integration

Run multiple models on a node with containerized inference server



Embedded | Automotive | Data center

Jetson | Drive | Tesla

developer.nvidia.com/tensorrt

# DEEPSTREAM SDK OVERVIEW

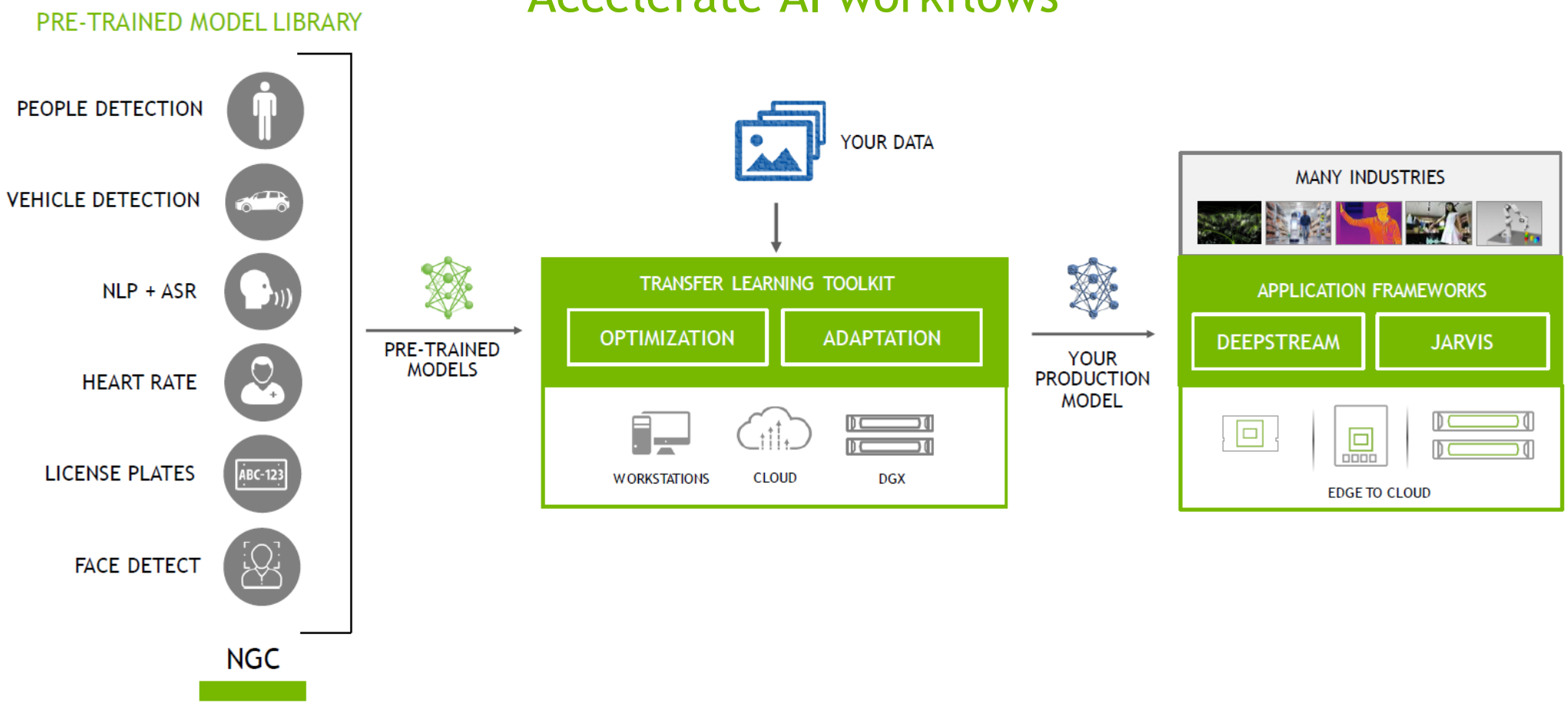
## Transform pixel and sensor data to actionable insights

- **Streaming analytics toolkit** for AI-based multi-sensor processing, video and image understanding with **TLS security**
- Deploy on the edge and connect to any cloud
- **C/C++ and python** choice of development
- Extensive AI model support: **SSD, YOLO, FasterRCNN, and MaskRCNN** and more
- **Flexibility** for rapid prototyping to full production
- **Speed up overall development efforts** by training with TLT and deploying with DS
- **Turnkey integration** with AWS IoT and Azure IoT
- Select from 15+ existing DS custom plugins or create your own

### Deploy Your Vision AI Application Anywhere

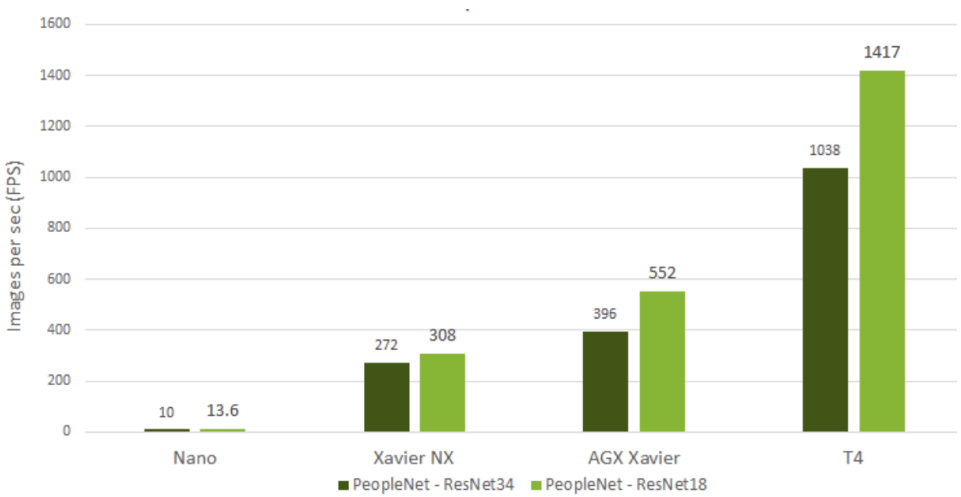
# TRANSFER LEARNING TOOLKIT

## Accelerate AI workflows

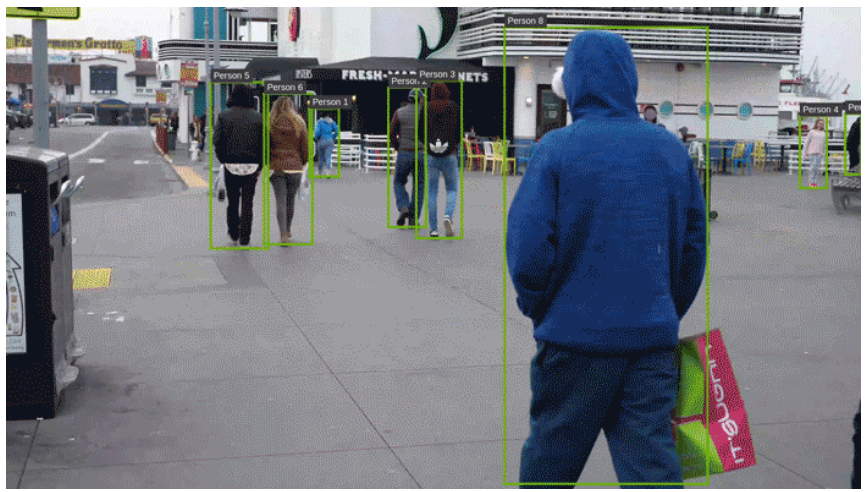


# PEOPLENET: REAL-TIME INFERENCE PERFORMANCE

## Detect persons, bags and faces




VIDEO DEMO



Number of classes: 3
Dataset: 750k frames

| Accuracy |
| --- |
| 84% |

https://ngc.nvidia.com/catalog/models/nvidia:tlt_peoplenet

# NVIDIA GAZE ESTIMATION PRE-TRAINED MODEL

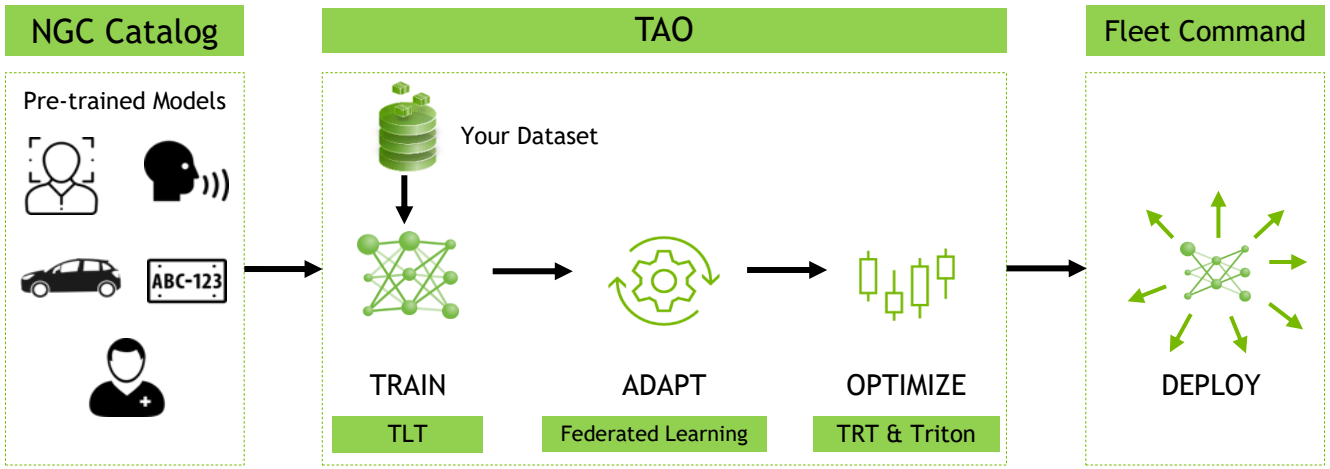
# VIDEO DEMO

## TrafficCamNet and VehicleMakeNet In Action


https://ngc.nvidia.com/catalog/models/nvidia:tlt_trafficcamnet
https://ngc.nvidia.com/catalog/models/nvidia:tlt_vehiclemakenet

# NVIDIA TAO FRAMEWORK

## Train | Adapt | Optimize




### TRAIN

UI based framework simplifies AI development

Domain specific models in hours v. months

### ADAPT & OPTIMIZE

Increase model accuracy with federated learning

Optimize with TensorRT

### DEPLOY

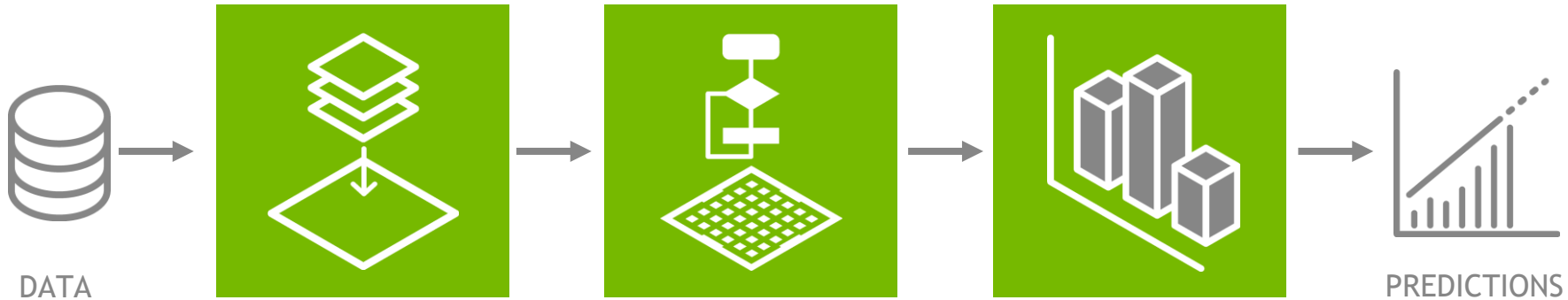
Deploy from anywhere to everywhere

Effortless management and secure deployments

NVIDIA TAO availability: 2H, 2021

# RAPIDS

## GPU-ACCELERATED DATA SCIENCE WORKFLOW




### DATA PREPARATION - ETL

Python drop-in **pandas** replacement built on CUDA C++.

GPU-accelerated Spark

### MODEL TRAINING

GPU-acceleration of today's most popular ML algorithms such as **XGBoost**

Easy-to-adopt, **scikit-learn** like interface

### VISUALIZATION

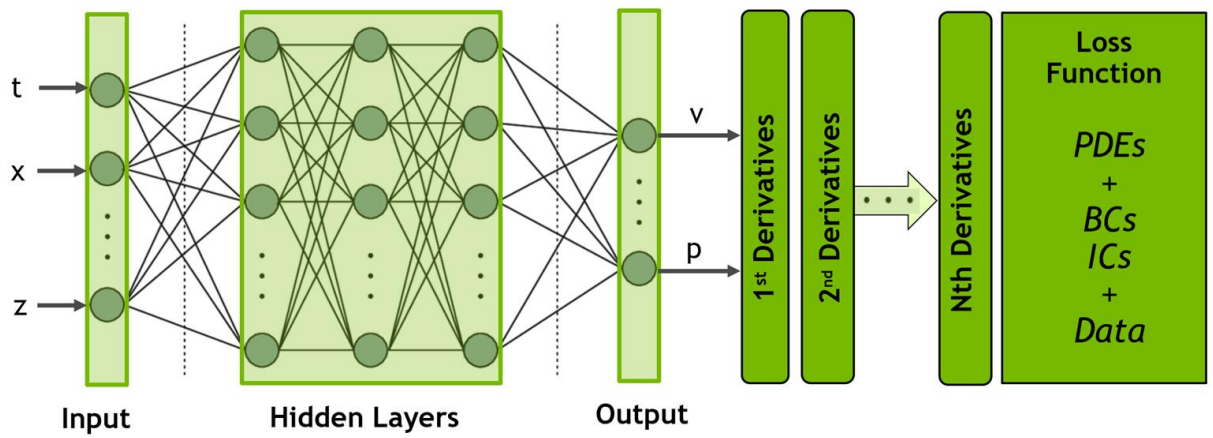
Effortless exploration of datasets, billions of records in milliseconds

Dynamic interaction with data = faster ML model development

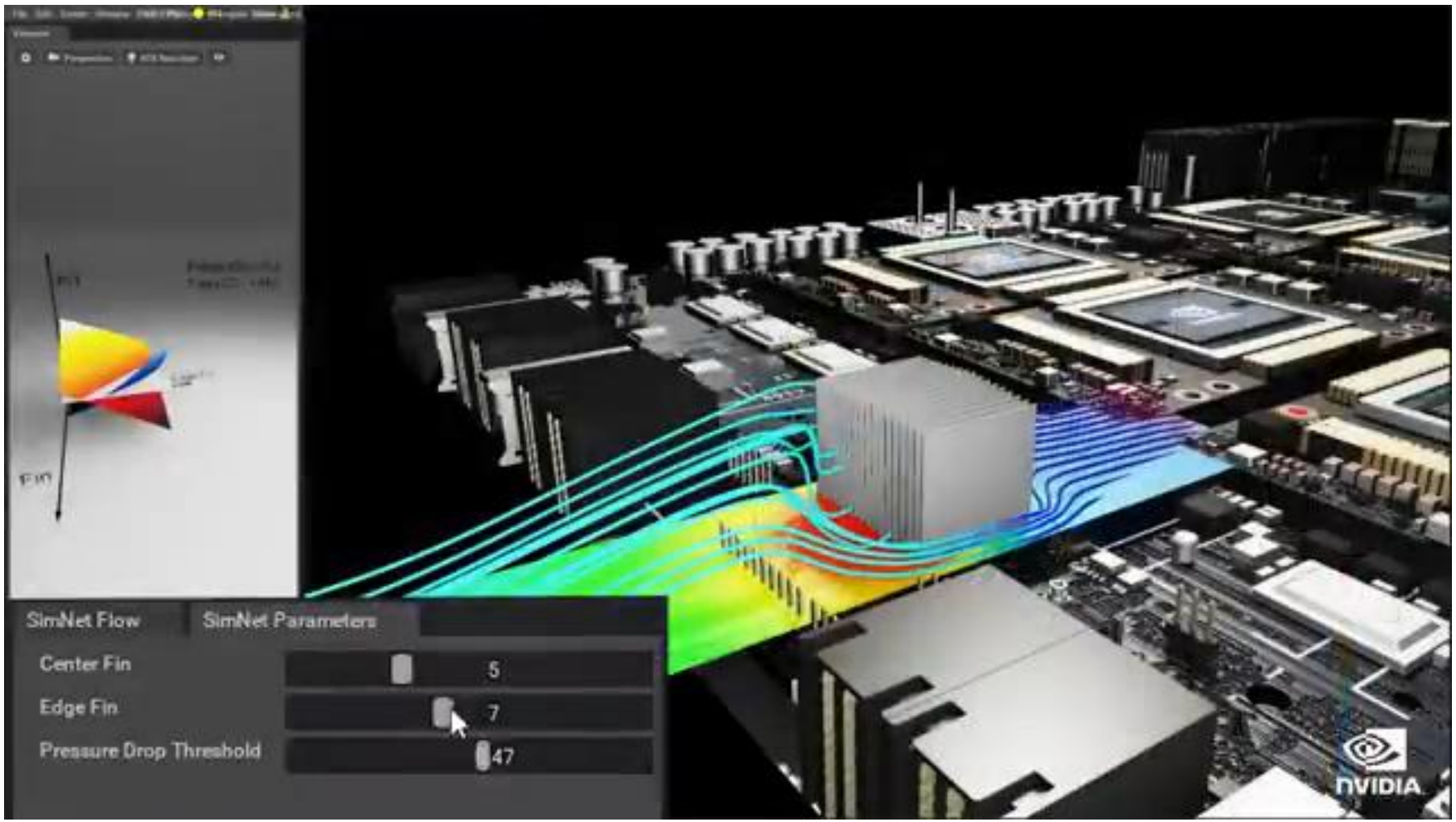
# SIMNET

## AI toolkit for multi-Physics Simulation

- General ODE/PDE neural network solver:
  - Engineering Physics ( + Computational Bio/Chem, Hi-energy Physics, Finance etc.)
  - Strong/Differential or Weak/Variational form
- Physics driven
- General Geometry/Shape modeling
- Multiple network architectures and features
- Parametrization of Geometry & Physics
- Performance optimized for single & multiple GPUs/Nodes
- APIs for customized development

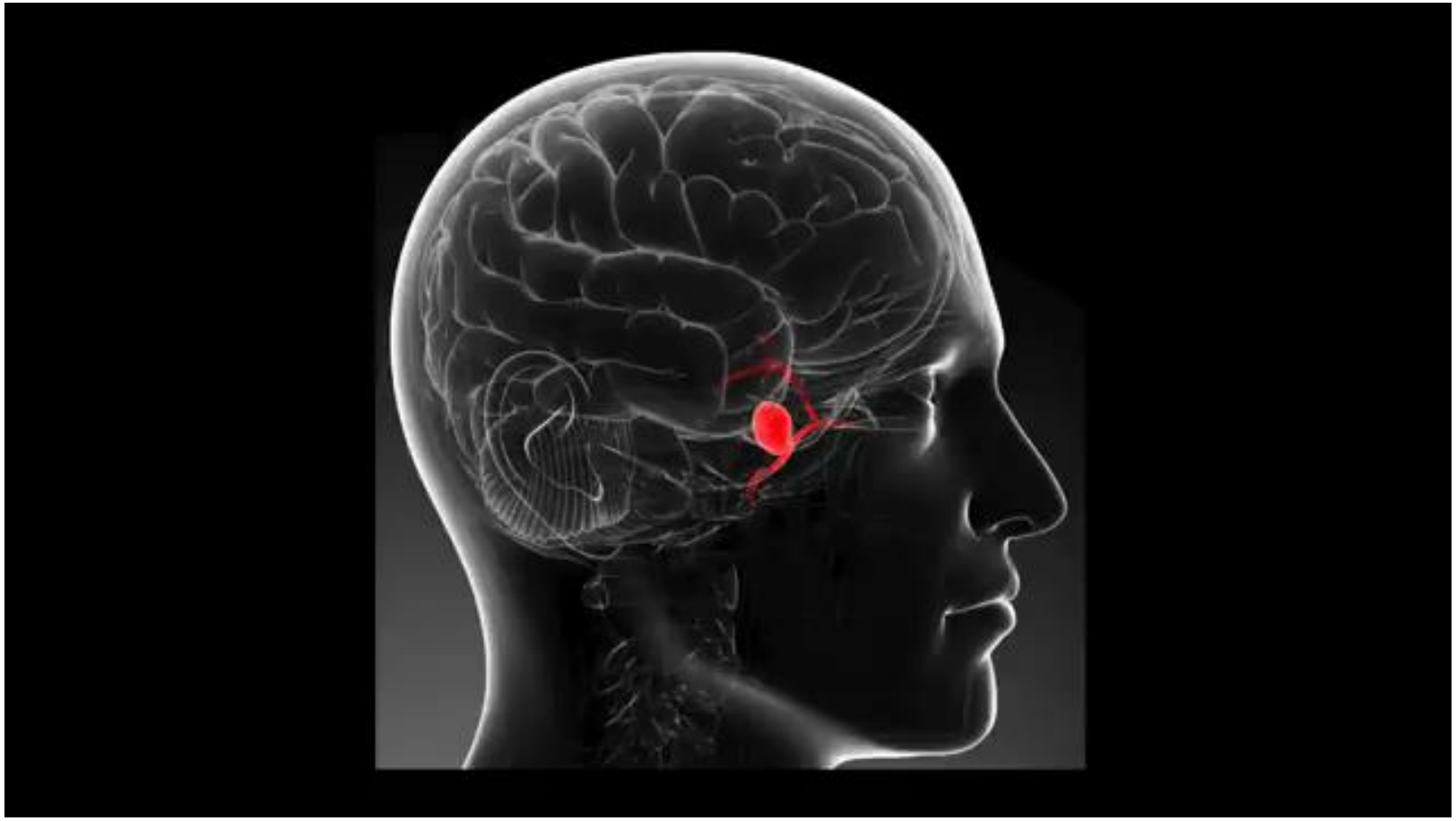


# SIMNET: OPTIMISING HEAT SINK DESIGN

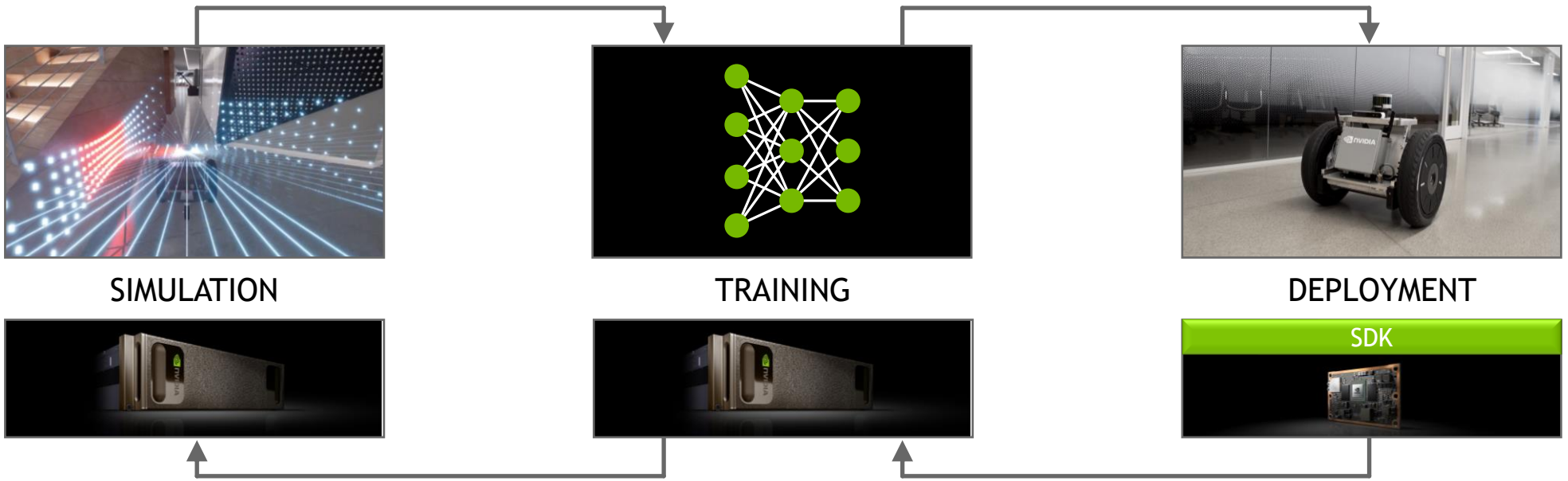


# FLOW PHYSICS QUANTISATION IN AN ANUERYSM

https://www.youtube.com/watch?v=QjY_8xFjsgE



# NVIDIA ISAAC ROBOTICS PLATFORM




*https://developer.nvidia.com/isaac-sdk*

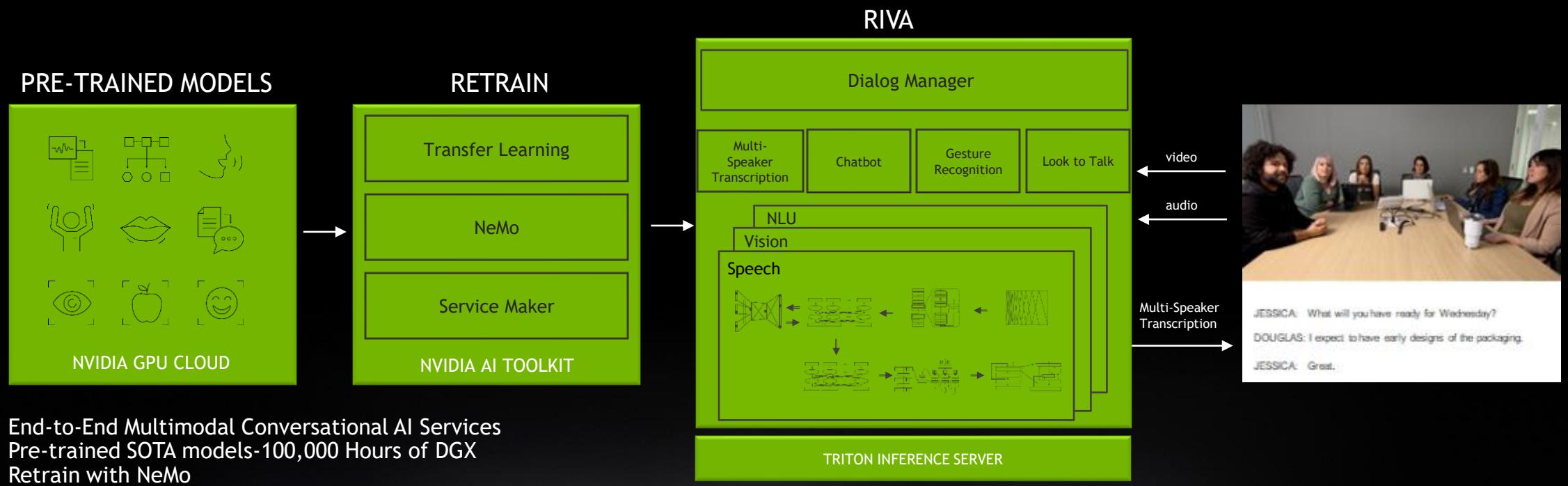
REINFORCEMENT LEARNING – ISAAC GYM

# ISAAC

## Virtual environment (omniverse)

# RIVA

## Framework for Multimodal Conversational AI services

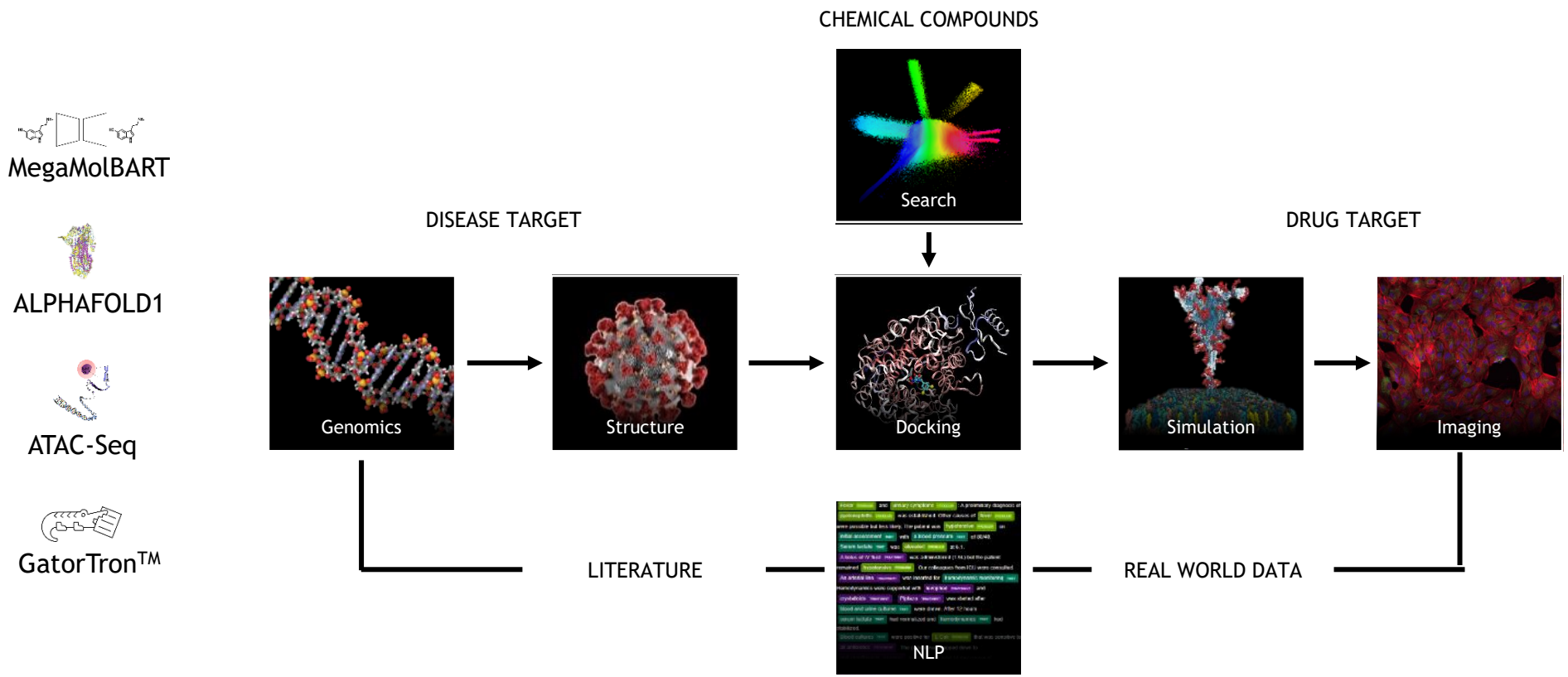



End-to-End Multimodal Conversational AI Services
Pre-trained SOTA models-100,000 Hours of DGX
Retrain with NeMo
Interactive Response – 150ms on A100 versus 25sec on CPU
Deploy Services with One Line of Code

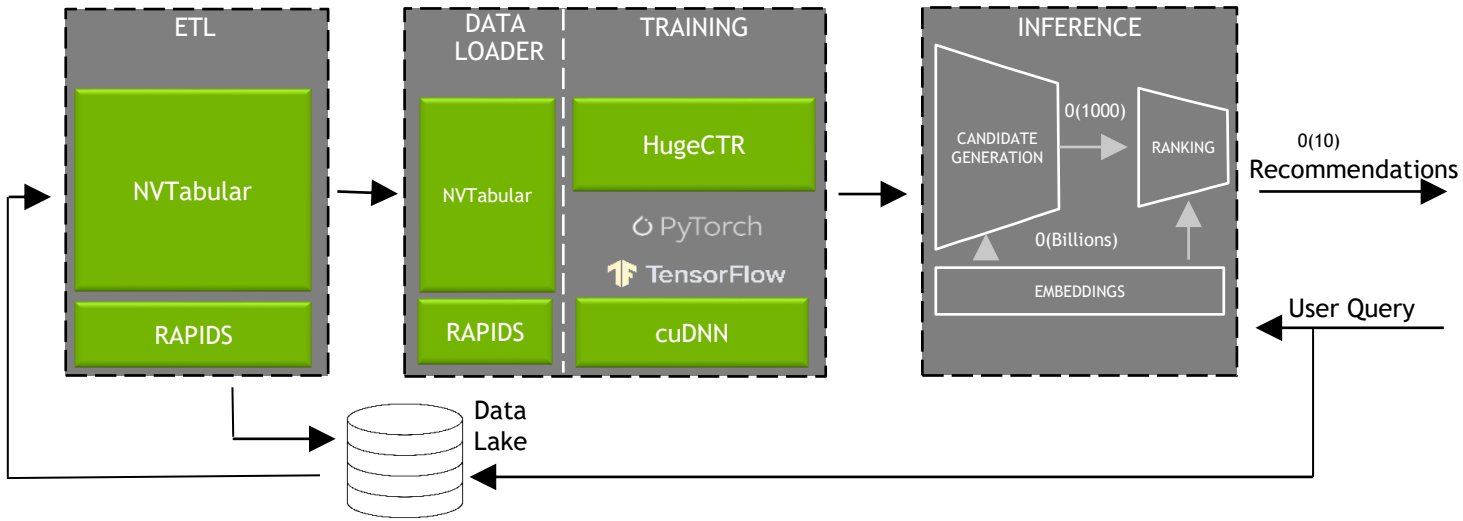
Sign-up for Early Access:
developer.nvidia.com/riva

# HEALTHCARE – CLARA DISCOVERY

## Understanding Disease and Discovering Therapies



# NVIDIA MERLIN END-TO-END ACCELERATED RECOMMENDER SYSTEM




## DRAMATICALLY LOWERED COSTS AND IMPROVED CONTENT RANKING LATENCY

- Snapchat leveraged NVIDIA T4 GPUs and Merlin
- Improved ML inference cost efficiency by 50%
- Decreased serving latency by over 60%
- Provided compute headroom to deploy heavier, more accurate ad and content ranking models

## TASTEFUL RECOMMENDATIONS FROM 600K RESTAURANTS

- 600,000 merchants and serve 80% of US Households.
- Restaurant and dish recommender adopted NVTabular from NVIDIA Merlin
- Reduced training time from 1 hour on CPU to 5 minutes on GPU
- Reduced cost by 95% on NVIDIA A100 GPUs

Postmates

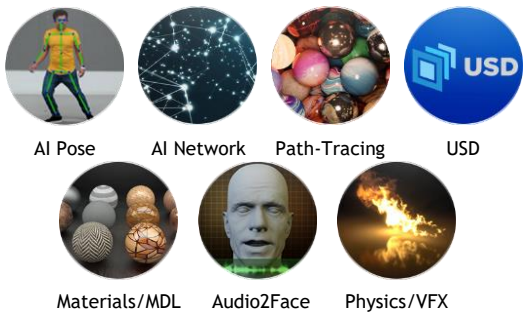
# A NEW ERA OF COLLABORATION AND SIMULATION

## CONNECTORS

Connectors for Blender, Adobe Substance, Autodesk Maya, Epic Games' Unreal Engine, Trimble SketchUp with more to come.

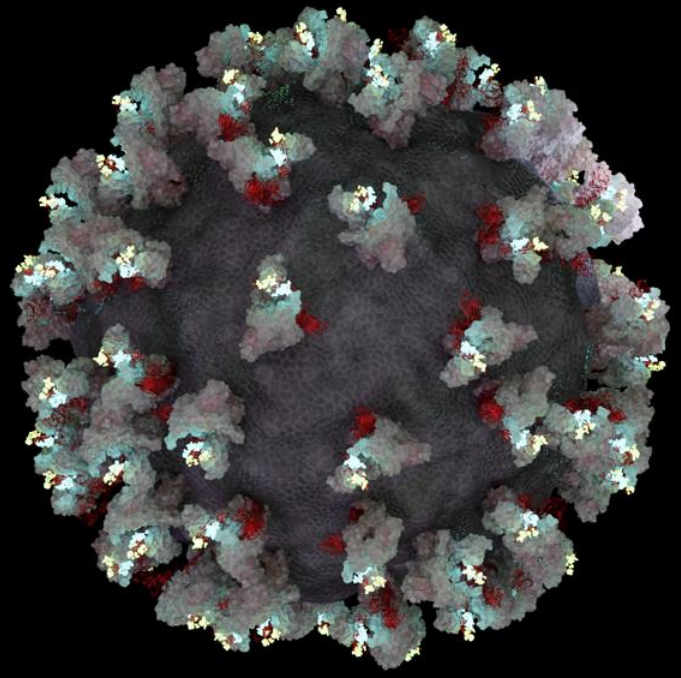

## PORTAL

## TECH

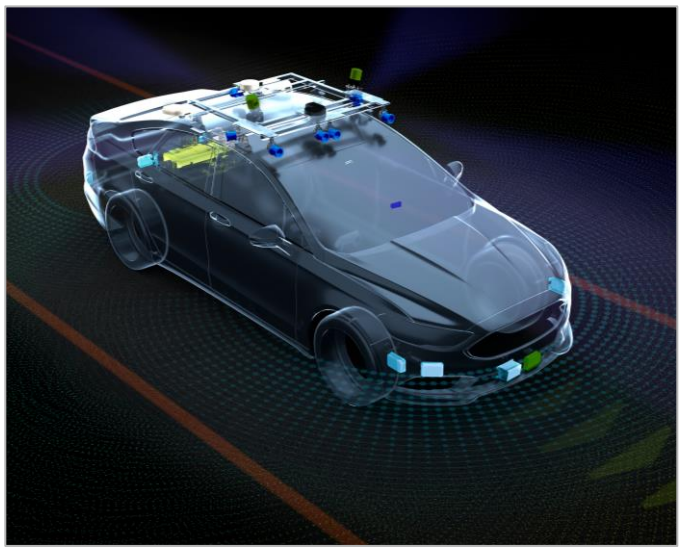


# OMNIVERSE

## Digital twin – BMW Factory



# AUTONOMOUS VEHICLE DEVELOPMENT AND VALIDATION



Highly Complex System
Large Computers, DNNs, Sensors

Real-Life Scenario Coverage
Account for Rare & Unpredictable Cases

Continuous Reaction Loop
Vehicle & World are Dependent


BRAKE
ACC
49 mph

# RESOURCES

# NGC: GPU-OPTIMIZED SOFTWARE HUB

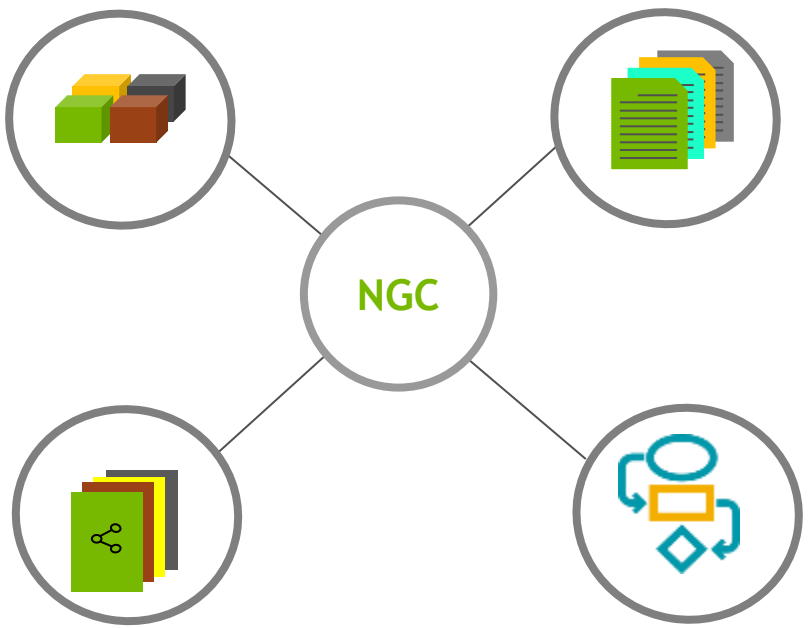
## Simplifying DL, ML and HPC Workflows

### 50+ Containers

DL, ML, HPC

### 15+ Model Training Scripts

NLP, Image Classification, Object Detection and more




### 60 Pre-trained Models

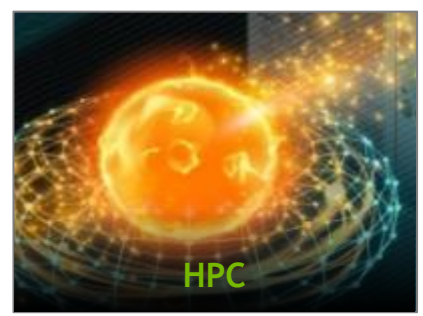
NLP, Image Classification, Object Detection and more

### Workflows

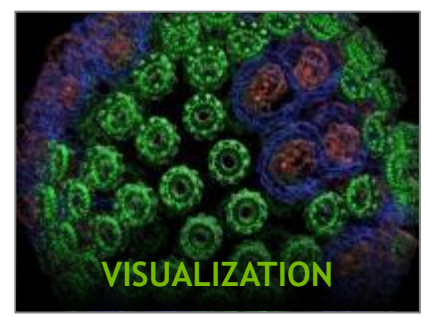
Medical Imaging, Intelligent Video Analytics


TensorFlow | PyTorch | more


RAPIDS | H2O | more




NAMD | GROMACS | more




ParaView | IndeX | more

NVIDIA

# DEVELOPER ENGAGEMENT PLATFORMS

| | |
|---|---|
| Information, downloads, special programs, code samples, and bug submission | developer.nvidia.com |
| Containers for cloud and workstation environments | ngc.nvidia.com |
| Insights & help from other developers and NVIDIA technical staff | devtalk.nvidia.com |
| Technical documentation | docs.nvidia.com |
| Deep Learning Institute: workshops & self-paced courses | courses.nvidia.com |
| In depth technical how to blogs | devblogs.nvidia.com |
| Developer focused news and articles | news.developer.nvidia.com |
| Webinars | nvidia.com/webinar-portal |
| GTC on-demand content | https://www.nvidia.com/on-demand/ |

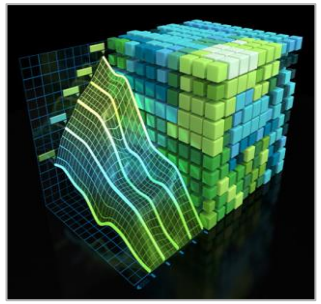
# DEEP LEARNING INSTITUTE (DLI)

Hands-on, self-paced and instructor-led training in deep learning and accelerated computing

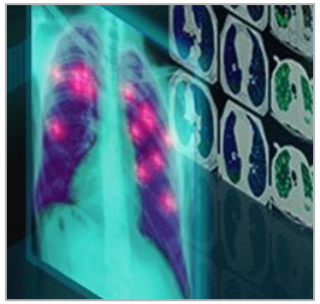
Request onsite instructor-led workshops at your organization:
www.nvidia.com/requestdli

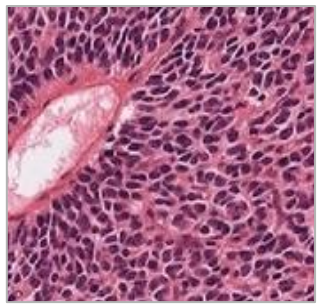
Take self-paced courses online:
www.nvidia.com/dlilabs

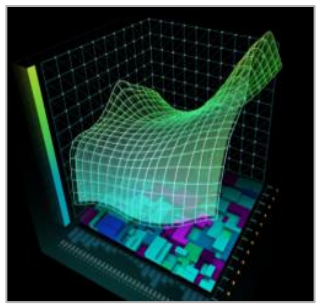
Download the course catalog, view upcoming workshops, and learn about the University Ambassador Program:
www.nvidia.com/dli


Accel. Computing Fundamentals


Autonomous Vehicles

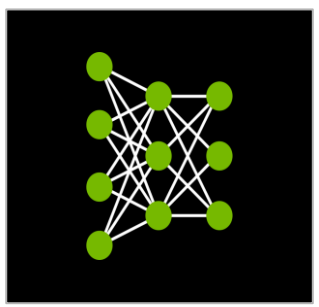

Medical Image Analysis


Genomics


Finance


Digital Content Creation


Game Development


Deep Learning Fundamentals

More industry-specific training coming soon…

# RESOURCES AVAILABLE TO ACADEMICS

**Developer Teaching Kits**: which include free access to online training for students but they have to be requested by a lecturer/professor.

**Academic Workshops**:
The NVIDIA website lists free academic workshops that our Ambassadors are giving around the world that you can go and attend

**Bootcamps**:
~ 2 day tailored training events, typically for a target group e.g. OpenACC, AI for Science

**Hackathons**:
In-depth events with access to NV *devtech*

NVIDIA GTC 21

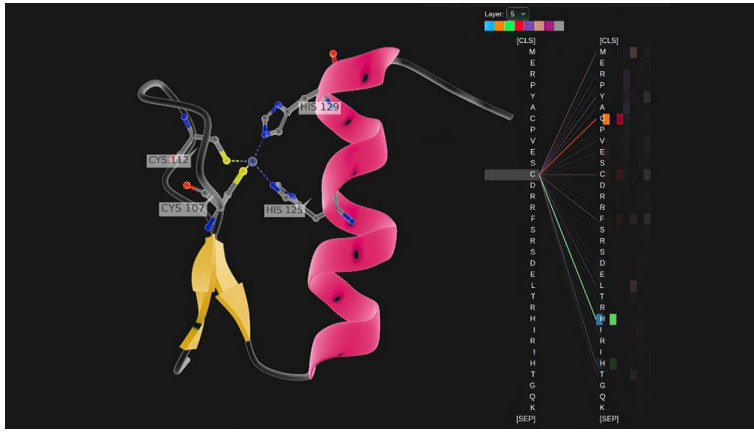
# THE CONFERENCE FOR AI INNOVATORS, TECHNOLOGISTS, AND CREATIVES

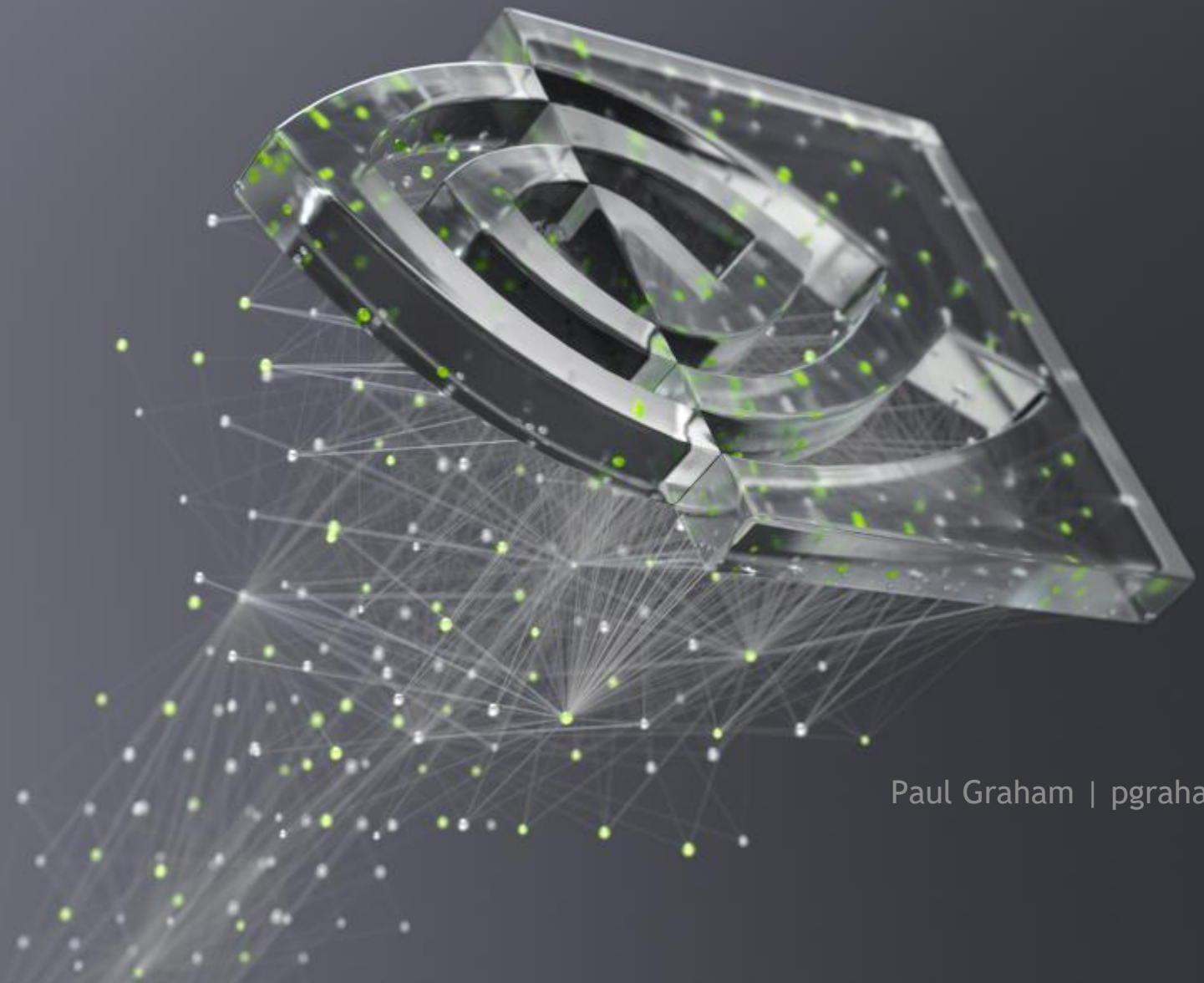
Join us at GTC Fall 2021 on Nov 8 – 11 for the latest in AI, HPC, healthcare, game development, networking, and more.

NVIDIA's GTC brings together a global community of developers, researchers, engineers, and innovators to experience global innovation and collaboration.

Don't miss out on the exclusive GTC keynote by Jensen Huang on **Nov 9**, available to everyone.

Visit https://www.nvidia.com/gtc to learn more and register for free



Paul Graham | pgraham@nvidia.com

NVIDIA